Hindawi Publishing Corporation
Advances in Decision Sciences
Volume 2015, Article ID 925935, 7 pages
http://dx.doi.org/10.1155/2015/925935

Research Article

A Framework for the Selection of Binarization Techniques on Palm Leaf Manuscripts Using Support Vector Machine

Rapeeporn Chamchong[1,2] and Chun Che Fung[1]

[1] *School of Engineering and Information Technology, Murdoch University, Perth, WA 6150, Australia*
[2] *Department of Computer Science, Faculty of Informatics, Mahasarakham University, Maha Sarakham 44150, Thailand*

Correspondence should be addressed to Rapeeporn Chamchong; rapeeporn.c@gmail.com

Received 3 September 2014; Revised 10 December 2014; Accepted 31 December 2014

Academic Editor: Henry Schellhorn

Challenges for text processing in ancient document images are mainly due to the high degree of variations in foreground and background. Image binarization is an image segmentation technique used to separate the image into text and background components. Although several techniques for binarizing text documents have been proposed, the performance of these techniques varies and depends on the image characteristics. Therefore, selecting binarization techniques can be a key idea to achieve improved results. This paper proposes a framework for selecting binarizing techniques of palm leaf manuscripts using Support Vector Machines (SVMs). The overall process is divided into three steps: (i) feature extraction: feature patterns are extracted from grayscale images based on global intensity, local contrast, and intensity; (ii) treatment of imbalanced data: imbalanced dataset is balanced by using Synthetic Minority Oversampling Technique as to improve the performance of prediction; and (iii) selection: SVM is applied in order to select the appropriate binarization techniques. The proposed framework has been evaluated with palm leaf manuscript images and benchmarking dataset from DIBCO series and compared the performance of prediction between imbalanced and balanced datasets. Experimental results showed that the proposed framework can be used as an integral part of an automatic selection process.

1. Introduction

Binarization of ancient document images is a crucial process to remove unrelated artefacts and background noise in document images. Image binarization is essential not only to document image analysis, but also to significantly improve the overall performance as poor binarization will result in poor recognition of the original characters and additional noise could be added to the image. Therefore, determining proper binarization techniques can be a key significant factor in achieving promising results from document image analysis.

There are many binarization techniques [1–4] that have been proposed in the literature. Some techniques performed well on certain datasets but not on the others. Due to this reason, a selection of binarization techniques can be a key step in improving the performance of document image analysis. In most circumstances, a human operator determines and compares results from different processing techniques and then selects one of the approaches based on visual impression, examination, or intuition. However, for an automated approach, there must be some forms of quantitative assessment so that the "optimal" technique is selected. If an automated selection process is implemented, this will assist and improve the system performance. This study therefore aims at proposing a selection process for the most appropriate binarization technique by machine learning, and in particular, the selection is based on *Support Vector Machines* (SVMs) due to its appropriateness for classification problems [5].

Over the past five centuries, palm leaves have been used as one of the most popular media for written documents in Asian regions. These ancient documents are heritage passed down through many generations. Libraries and museums all across Thailand contain a large collection of palm leaf manuscripts written in ancient local languages. Currently, scanners are able to binarize documents with a good contrast of foreground components and a uniform background [6].

However, most of the palm leaf manuscripts are of poor quality due to smeared or smudged characters, poor writing, and nonuniform changes in colors due to long term storage.

In this study, a proposed framework has been applied to select the binarization techniques with practical dataset that was collected by the project for Palm Leaf Preservation in Northeastern Thailand Division, Mahasarakham University [7]. The benchmarking dataset from DIBCO series [8–13] was also used to evaluate the framework. In this research, binarization techniques for degradation document have been used in the proposed method that are *Adaptive Logical Level* (ALL) technique [4], *Improvement of Integrated Function* (IIF) algorithm [3], *Background Estimation* (BE) technique [14], and *Local Maximum and Minimum* (LMM) technique [2]. It first extracts feature patterns from a grayscale image by considering global intensity, local contrast, and intensity. As the data in the datasets could be imbalanced, Synthetic Minority Oversampling Technique (SMOTE) [15] is used to synthesize the data in order to provide the balance and improve the performance.

The remaining of this paper is structured as follows: a proposed selection framework of binarization techniques is described in the next section. In Section 3, experimental results are then given and, finally, followed by a conclusion of this work.

2. Selection Framework of Binarization Techniques

This section explains the selection process based on machine learning technique. The selection is performed by classifying the appropriate techniques based on the features extracted from the image. In this study, the issue of imbalanced data has been addressed in order to improve the accuracy. SVM is then used to select the appropriate binarization technique for generating the binary image. Figure 1 illustrates the overall process of the proposed method for selecting the optimal technique. The dataset is first separated into two sets, that is, training and test set for the learning process. To select an optimal binarization technique, feature extraction is performed initially. Objects in the images can be extracted and analyzed using generic image processing techniques. This work used a real word dataset in which the data class is not unevenly distributed, leading to the imbalanced problem for multiclass classification. The data is then normalized. For these features, there are various feature subsets, which are represented from different perspectives from the global and local properties of the images. To select the promising discriminative features, feature selection was performed. The feature patterns are then classified in order to predict the appropriate binarization technique.

2.1. Feature Extraction. Feature extraction is an essential step in any learning method which transforms the characteristics of original data to feature patterns for decision making. This subsection explains the feature pattern of the images used in the dataset, and *Principal Component Analysis* (PCA) is used for dimensionality reduction of the feature space.

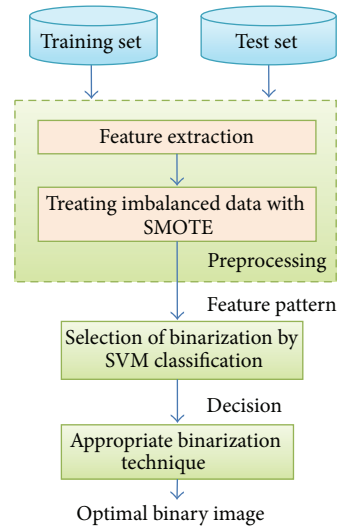

Figure 1: Overall process of the proposed method for selecting the optimal binarization techniques.

2.1.1. Feature Pattern. The most commonly used features applied to global binarization techniques are intensity histograms. They are used to convey the colour distribution information. This forms a compact representation of the colour feature. Furthermore, the mean, standard deviation, minimum, and maximum of intensity are also used as global features.

For local binarization techniques, intensity and contrast have been the most frequently used features [16]. A contrast feature has also been used and modified by Su et al. [17]. If a significant intensity change occurs at the boundary of the foreground text and the background, the contrast of grayscale indicates the characteristics differentiation between the foreground and background. In this study, the contrast feature has been used for feature extraction. The contrast value of this study has been modified by decomposing the image into subimages. In addition, this study also applied the intensity values by using mean, standard deviation, maximum, and minimum of intensity of the subareas. The features of an image used in this study are explained below.

(a) Global Features. Histogram of an image represents the relative frequency of occurrence of the various gray levels in the image. It gives a global description of the image and the shape of the histogram reveals significant contrast information. A discrete function of the histogram [18], H, is given by the relation

$$H = \{h_0, h_1, \ldots, h_{L-1}\}, \tag{1}$$

where

$$h_l = \frac{n_l}{N}, \tag{2}$$

while l is the level of grayscale that $0 \le l \le L-1$, n_l is the number of pixels in the image with the lth level of grayscale, and N is the total number of pixels in the image.

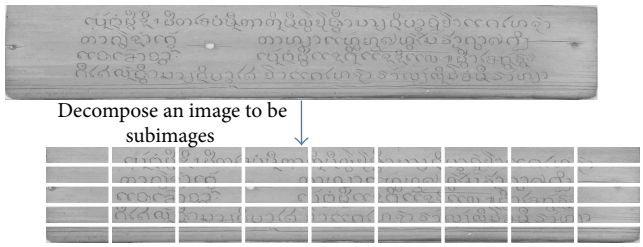

Figure 2: Image decomposition.

The image histogram carries important content of the image. For global binarization techniques based on clustering, this content is useful to distinguish two objects between foreground and background of an image.

In this study, 64 bins of grayscale histogram of the image are extracted and used as features for the selection module. This represents the global characteristic of the image and it could be used to assist the decision on selecting the appropriate technique.

The mean and standard deviation of the intensity of an image [19] represent the compact features. The mean of an image (f_μ) captures the first-order moment, and the standard deviation of the image (f_σ) is captured as the second-order moment. These expressions are shown as follows:

$$f_\mu = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N} f(x,y),$$
$$f_\sigma = \sqrt{\frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}\left(f(x,y)-f_\mu\right)^2}, \quad (3)$$

where $f(x, y)$ is the intensity value of the colour pixel at (x, y) axis, while M is the number of columns and N is the number of rows of image.

The other two intensity features are minimum $f_{\min}$ and maximum $f_{\max}(i, j)$ intensity values of image which were also used in this study.

(b) Local Features. For image binarization, intensity and contrast have been widely used as descriptors to classify foreground and unrelated objects in images [16]. An example of image binarization implemented using contrast features was proposed by Su et al. [17]. As a significant intensity change at the boundary of the foreground text and the background, the contrast of gray level relates to the characteristics of foreground and background. In this paper, contrast features are used as a local feature by decomposing the image into subimages as shown in Figure 2. In addition, this study also uses pixel intensity-based features by calculating the mean, standard deviation, maximum, and minimum of intensity of the subimages.

The contrast feature in local neighborhood from Su et al.'s study [17] is defined as follows:

$$\mathrm{Cont}(x,y) = \frac{f_{\max}(x,y) - I(x,y)}{f_{\max}(x,y) + \varepsilon}, \quad (4)$$

where $I(x, y)$ and $f_{\max}(x, y)$ denote the intensity of pixel (x, y) and the maximum intensity values within local area. ε is a positive value but infinitely small number, which is added in case the local maximum is equal to 0. Cont(x, y) refers to the contrast value of the estimating pixel (x, y).

In this study, the contrast feature is calculated in each subimage, and this feature is then modified as the following expression:

$$\mathrm{Cont}(i,j) = \frac{f_{\max}(i,j) - f_\mu(i,j)}{f_{\max}(i,j) + \varepsilon}, \quad (5)$$

where $f_\mu(i, j)$ and $f_{\max}(i, j)$ denote the average intensity of subimage (i, j) and the maximum intensity values of subimage (i, j). ε is a positive value but infinitely small number, which is added in case the maximum intensity values are equal to 0. Cont(i, j) refers to the contrast value of the estimating subimage (i, j), and $1 \le i \le C$ and $1 \le j \le R$ where C and R are the numbers of columns and rows of decomposed image.

The mean and standard deviation of the intensity of an image [19] represent the compact features. The mean of an image (f_μ) captures the first-order moment, and the standard deviation of the image (f_σ) is captured as the second-order moment. These expressions are shown as follows:

$$f_\mu = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N} f(x,y),$$
$$f_\sigma = \sqrt{\frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}\left(f(x,y)-f_\mu\right)^2}, \quad (6)$$

where $f(x, y)$ is the intensity value of the colour pixel at (x, y) axis, while M is the number of columns and N is the number of rows of the image.

The other two intensity features are the minimum $f_{\min}(i, j)$ and maximum $f_{\max}(i, j)$ intensity values of subimage which were used in this study.

Forty-five subimages have been considered in this study. 225 feature patterns from five features (contrast, mean, standard deviation, maximum, and minimum) are then extracted from an image. The number of overall features is 292 including 69 feature patterns of global features.

In machine learning, it is often unavoidable to have data with high dimensionality input features. To improve the prediction performance and yet provide faster and low cost predictors, dimensionality reduction can be applied. *Principal Components Analysis* (PCA) [20] is the most widely used technique and powerful tool for feature selection in the transformed space to reduce dimension feature. This technique is an unsupervised method based on a correlation or covariance matrix that has found application such as face recognition and image compression.

2.1.2. Principal Component Analysis. In machine learning, it is common to deal with data having high dimensionality input features. To improve the prediction performance, dimensionality reduction may be applied through a transformation of the original data.

I is the input dataset
M is the set of minority class instances
For each instance x_i in M
  Find the k-Nearest Neighbours (minority class instances) to x_i in M
  Obtain $\widehat{x}_i$ by randomising one from k instances
  δ = random number between zero and one
  $x_{\text{new}} = x_i + (\widehat{x}_i - x_i) \times \delta$
  Add x_{new} to I
End for

Algorithm 1: SMOTE algorithm [15].

Principal Components Analysis (PCA) is the most widely used technique [21] and a powerful tool [20] for feature selection in the transformed space for dimension reduction. This technique is an unsupervised method based on a correlation or covariance matrix that has been used in applications such as face recognition and image compression [20].

PCA [22] is calculated from the eigenvectors and eigenvalues of the data covariance matrix. The process is to find the axis system where the covariance matrix is diagonal. This technique can reduce the dimension of the representation. On the other hand, the original information content will be preserved as much as possible. The next subsection addresses the problem of imbalanced data for machine learning.

2.2. Treating Imbalanced Data with SMOTE. Real world datasets usually have the problem of imbalanced data. It is a significant problem affecting learning algorithms. It associates with the situations that some classes have much larger numbers of instances than the others. Examples of real world cases with an imbalanced dataset are biomedical applications, fraud detection, and network intrusion [23].

The issue of imbalanced data needs to be approached at data level with the objective of balancing the training data before learning process is applied. Approaches to deal with imbalanced data can be separated into three categories: undersampling, oversampling, and combined techniques. The undersampling technique aims to balance the dataset by removing instances of majority class while oversampling aims to balance the dataset by adding the minority class. In addition, the combined technique is a combination of both undersampling and oversampling techniques.

In this study, there are cases that contain only a few instances in the dataset; oversampling technique is therefore adopted. There are several oversampling techniques such as the *Random Oversampling Technique* and *Synthetic Minority Oversampling Technique* (SMOTE) [15]. SMOTE has been shown to be a successful method in many applications [23] and the SMOTE algorithm generates synthetic data based on the feature space similarities between minority examples. Other techniques such as *Random Oversampling Technique* perform oversampling by replicating minority class instances randomly. For this reason, the SMOTE algorithm may avoid the overfitting problem [24] and, in this study, the SMOTE algorithm used is shown in Algorithm 1.

The number of new minority class instances is increased by the above algorithm and the synthetic instances are generated by *Euclidian distance* technique. The minority class instances that are close together are considered first, before they are employed to form new minority class instances.

In this study, the class imbalanced problem in multiclass data was addressed with the *One-Against-All* (OAA) scheme [25]. The OAA scheme is a promising technique of multiclass problem and is suitable for small sized training data [26]. This scheme can be used to deal with the data balancing issue in multiclasses and it can also reduce the complexity in the machine learning process. As some of the datasets in this study contain fewer instances, the OAA scheme was therefore applied.

2.3. Selection Module. A selection process of this study is based on *Support Vector Machine* (SVM) due to its appropriateness for classification problems. SVM is a kind of classification based on statistical learning theory which was introduced by Vapnik [27], and its applications have provided good results. In this study, SVM was used to select the appropriate binarization technique by learning from feature patterns of a training dataset. The binarization technique is then used to generate the binary image.

The decision function of SVM is calculated from a training dataset. In this study, radial basis functions (RBFs) were used to separate the classes. For building this module, the libSVM library [28] has been used in the implementation.

3. Experimental Results

In this experiment, the framework of the selection was evaluated by using SVM. In the next subsection, the selection framework is evaluated with the real world dataset of palm leaf manuscript and the dataset of DIBCO series including DIBCO 2009, H-DIBCO 2010, DIBCO 2011, H-DIBCO 2012, and DIBCO 2013 [8–13]. The ground truth of palm leaf images was selected by visual human while the ground truth images of DIBCO series were generated following a semiautomatic procedure based on [29]. In the following subsection, the original binarization techniques and the proposed framework were compared with the dataset of DIBCO series.

Table 1: Performance of the selection of binarization techniques using SVM on imbalanced and balanced dataset.

Measure	Class	Dataset from DIBCO series			Palm leaf dataset (1 : 2 : 2)	
		Imbalanced dataset	Balanced dataset 1 by SMOTE (1 : 2 : 2)	Balanced dataset 2 by SMOTE (2 : 4 : 4)	Imbalanced dataset	Balanced dataset by SMOTE
F-measure	ALL	0.000	0.370	0.632	0.231	0.865
	LMM	0.778	0.650	0.679	0.735	0.857
	BE	0.000	0.615	0.882	0.033	0.943
	IIF	0.000	0.522	0.889	0.000	0.949
Accuracy		0.636	0.592	0.754	0.588	0.980
G-mean		0.000	0.448	0.736	0.000	0.902
AUC		0.381	0.828	0.949	0.683	0.980

3.1. Evaluation of the Selection Framework. In general, accuracy is used to illustrate the overall classification performance. In case of imbalanced dataset, it is premised that if the number of prior classes is very different, this measure may be unsuitable because misclassification may occur [24]. Other evaluation measures of the imbalanced problem have been proposed and they are *F-measure* (FM), *the geometric mean* (G-mean), and *the area under the ROC curve* (AUC) [23, 30]. These indicators aim to maximize the accuracy between the minority class and the majority class so they are good for the class imbalanced problem. These measures were therefore applied to evaluate the performance of the selection of the binarization techniques in this study.

The overall features comprised 68 global features (64 bins of histogram, a minimum, a maximum, a mean, and a standard deviation value of intensity of image) and 225 local features (45 subimages of contrast, 45 subimages of mean, 45 subimages of standard deviation, 45 subimages of maximum, and 45 subimages of minimum). In this experiment, benchmarking dataset and palm leaf manuscript dataset were used. In this experiment, 10-fold cross-validation was used in each dataset. The datasets of the experiment were separated into two types which are imbalanced dataset and balanced dataset (applied imbalanced dataset by SMOTE). The details of each dataset are described as the following subsection.

3.1.1. Benchmarking Dataset from DIBCO Series. This dataset was divided into three categories that are the following.

(1) Imbalanced Dataset. The dataset is composed of 66 instances (images), and there are four classes which are LMM 42 instances, ALL 7 instances, BE 12 instances, and IIF 5 instances (ratio of instances, LMM : ALL : BE : IIF = 64 : 10 : 18 : 7).

(2) Balanced Dataset 1. As class distribution of this dataset is imbalanced, SMOTE was applied to synthesize the minority classes. LMM is a majority class and other minority classes are ALL, BE, and IIF. The number of instances of minority classes in ALL was increased by 200 percent, in BE by 100 percent, and in IIF by 200 percent. The number of instances after synthesis was 103 instances, with 42 instances in LMM, 24 instances in ALL, 21 instances in BE, and 16 instances in IIF (ratio of instances, LMM : ALL : BE : IIF = 40 : 23 : 20 : 15).

(3) Balanced Dataset 2. This also applied the SMOTE to synthesize the minority classes from imbalanced dataset, but the number of instances of minority classes in ALL was increased by 400 percent, in BE by 200 percent, and in IIF by 400 percent. The number of instances after synthesis was 103 instances, with 42 instances in LMM, 36 instances in ALL, 35 instances in BE, and 25 instances in IIF (ratio of instances, LMM : ALL : BE : IIF = 30 : 26 : 25 : 18).

3.1.2. Palm Leaf Manuscript Dataset. The datasets of the experiment were separated into two groups that are the following.

(1) Imbalanced Dataset. The dataset is composed of 480 instances; there are four classes which are LMM 280 instances, ALL 96 instances, BE 58 instances, and IIF 46 instances (ratio of instances, LMM : ALL : BE : IIF = 58 : 20 : 12 : 10).

(2) Balanced Dataset. As class distribution of this dataset is imbalanced, SMOTE was applied to synthesize the minority classes. LMM is a majority class and other minority classes are ALL, BE, and IIF. The number of instances of minority classes in ALL was increased by 100 percent, in BE by 200 percent, and in IIF by 200 percent. The number of instances after synthesis was 784 instances, with 280 instances in LMM, 192 instances in ALL, 174 instances in BE, and 138 instances in IIF (ratio of instances, LMM : ALL : BE : IIF = 36 : 24 : 22 : 18).

The performance of selection of binarization techniques on imbalanced and balanced dataset is shown in Table 1.

With respect to the selection of imbalanced dataset, the performances of class LMM are significantly better than those from classes ALL, BE, and IIF, which have smaller instances. By applying SMOTE to balanced dataset, the performance of class LMM increased slightly, while the performances of classes ALL, BE, and IIF improved greatly. By applying SMOTE to the balanced dataset, the selection accuracy, G-mean, and AUC of imbalanced dataset are significantly improved in both datasets. In palm leaf dataset, performance of those terms improved by 40.8%; 90.2%; and 29.7%, respectively. In benchmarking dataset of balanced dataset group 2 of those terms also improved by 11.8%, 73.6%, and 56.8%, respectively. In benchmarking dataset of balanced dataset group 1 improved only in G-mean and AUC measures while accuracy was slightly decreased.

(a) Original image 1 (b) Binary result from image 1 (by LMM technique)

(c) Original image 2 (d) Binary result from image 2 (by ALL technique)

Figure 3: Two sample results from an automatic selection of binarization technique.

Table 2: Evaluation results of original binarization techniques and the proposed framework on benchmarking dataset.

Measures	Binarization techniques				
	ALL	IIF	BE	LMM	The proposed framework
F-measure	80.7873	55.2662	86.2899	89.7139	91.2494
PSNR	16.0682	14.6766	17.8789	19.1531	19.6587
DRD	8.8728	11.8249	8.2245	3.7656	2.8869
MPM	7.8459	1.3269	11.3429	2.5617	1.1085

3.2. A Comparison of the Proposed Framework and Original Binarization Techniques. In this study, the proposed technique was compared to the original binarization techniques including ALL, IIF, LMM, and BE with the dataset from DIBCO series. The evaluation measures of this study are described in [11] to compare the binarization images with the ground truth images. These measures consist of *F*-measure, PSNR, distance reciprocal distortion metric (DRD), and misclassification penalty metric (MPM). The evaluation result is shown in Table 2.

Based on the four original binarization techniques in Table 2, the LMM technique provided the best results in terms of *F*-measure, PSNR, and DRD while the IIF gave the best result in terms of MPM. Considering the proposed framework, it has superior results in all terms of measures than each of the original binarization techniques.

An automatic selection of multiple binarization techniques used in this framework has been applied to the users in recommending the appropriate technique. Figure 3 shows two sample results of an automatic selection from three binarization techniques (ALL, BE, and LMM) on palm leaf manuscripts.

4. Conclusions

This paper described experiments and results on document degradation and proposed a framework for an automatic selection from multiple binarization techniques by using SVM with imbalanced and balanced datasets (by applying SMOTE) for palm leaf manuscripts. The evaluation result was also evaluated on benchmarking dataset. The proposed measurement of learning for the selection is based on *F*-measure, accuracy, G-mean, and AUC. Another experiment, the original binarization techniques, and the proposed framework were compared based on *F*-measure, PSNR, DRD, and MPM.

With regard to the key points in this study, the automatic selection of binarization techniques used in this framework will be helpful to the users in recommending the appropriate technique. A comparison of imbalanced and balanced datasets of the selection framework in all terms of measures indicates that the selection works with better performance on balanced dataset than imbalanced dataset. However, the performance of the selection framework on balanced dataset by SMOTE is quite low in some measures if there are a few instances such as in benchmarking dataset group 1. Because of this, percent of minority classes must increase. This framework forms a prototype system for research in this area. It is noted that this framework still needs to be refined and the selection could be ranked by the user and modified as a semiautomatic approach. In addition, the top two or three rankings of the selection may also be combined as an integral of these techniques. Furthermore, there is a need to generate benchmarking datasets on palm leaf manuscripts for future research.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors wish to express their thanks and appreciation to the members of the Preservation of Palm Leaf Manuscripts Project, Mahasarakham University, Thailand, for their support and for providing images from their database. In particular, the authors thank Dr. Phatthanaphong Chomphuwiset for proofreading this paper.

References

[1] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, 2000.

[2] B. Su, S. Lu, and C. L. Tan, "Binarization of historical document images using the local maximum and minimum," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pp. 159–166, ACM, Boston, Mass, USA, 2010.

[3] Ø. Due Trier and T. Taxt, "Improvement of 'integrated function algorithm' for binarization of document images," *Pattern Recognition Letters*, vol. 16, no. 3, pp. 277–283, 1995.

[4] Y. Yang and H. Yan, "An adaptive logical method for binarization of degraded document images," *Pattern Recognition*, vol. 33, no. 5, pp. 787–807, 2000.

[5] F. van der Heijden, R. P. Duin, D. de Ridder, and D. M. Tax, *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*, John Wiley & Sons, West Sussex, UK, 2004.

[6] L. O'Gorman and R. Kasturi, *Document Image Analysis*, IEEE Computer Society Press, Los Alamitos, Calif, USA, 1995.

[7] Project for Palm Leaf Preservation in Northeastern Thailand Division, http://www.bl.msu.ac.th/2553/english_bl.htm.

[8] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 Document Image Binarization Contest (DIBCO 2009)," in *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR '09)*, pp. 1375–1382, Barcelona, Spain, July 2009.

[9] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "DIBCO 2009: document image binarization contest," *International Journal on Document Analysis and Recognition*, vol. 14, no. 1, pp. 35–44, 2011.

[10] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-DIBCO 2010—handwritten document image binarization competition," in *Proceedings of the International Conference on Frontiers in Handwriting Recognition (ICFHR '10)*, pp. 727–732, IEEE, Kolkata, India, November 2010.

[11] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 document image binarization contest (DIBCO 2011)," in *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR '11)*, pp. 1506–1510, Beijing, China, September 2011.

[12] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012)," in *Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR '12)*, pp. 817–822, Bari, Italy, September 2012.

[13] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2013 document image binarization contest (DIBCO 2013)," in *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR '13)*, pp. 1471–1476, August 2013.

[14] S. Lu, B. Su, and C. L. Tan, "Document image binarization using background estimation and stroke edges," *International Journal on Document Analysis and Recognition*, vol. 13, no. 4, pp. 303–314, 2010.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[16] M. Valizadeh and E. Kabir, "Binarization of degraded document image based on feature space partitioning and classification," *International Journal on Document Analysis and Recognition*, vol. 15, no. 1, pp. 57–69, 2012.

[17] B. Su, S. Lu, and C. L. Tan, "Combination of document image binarization techniques," in *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR '11)*, pp. 22–26, Beijing, China, September 2011.

[18] M. Cheriet, N. Kharma, C.-L. Liu, and C. Suen, *Character Recognition Systems : A Guide for Students and Practioners*, John Wiley & Sons, Hoboken, NJ, USA, 2007.

[19] T. Acharya and A. K. Ray, *Image Processing: Principles and Applications*, John Wiley & Sons, Hoboken, NJ, USA, 2005.

[20] L. I. Smith, *A Tutorial on Principal Components Analysis*, vol. 51, Cornell University, 2002.

[21] A. Webb, *Statistical Pattern Recognition*, John Wiley & Sons, London, UK, 2nd edition, 2002.

[22] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, USA, 2002.

[23] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[24] P. Jeatrakul, *Enhancing classification performance over noise and imbalanced data problems [Ph.D. thesis]*, School of Information Technology, Murdoch University, Perth, Australia, 2012.

[25] M. Aly, *Survey on Multi-Class Classification Methods* , California Institute of Technology, Pasadena, Calif, USA, 2005.

[26] P. Jeatrakul and K.-W. Wong, "Enhancing classification performance of multi-class imbalanced data using the OAA-DB algorithm," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '12)*, pp. 1–8, IEEE, Brisbane, Australia, June 2012.

[27] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[28] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[29] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "An objective evaluation methodology for document image binarization techniques," in *Proceedings of the 8th IAPR International Workshop on Document Analysis Systems (DAS '08)*, pp. 217–224, Nara, Japan, September 2008.

[30] Y. Sun, M. S. Kamel, and Y. Wang, "Boosting for learning multiple classes with imbalances class distribution," in *Proceedings of the 6th International Conference on Data Mining (ICDM '06)*, pp. 592–602, Hong Kong, December 2006.